1 Parameter tuning

We tune the hyper parameters using a 3 fold cross-validation (CV) on the training split. For all baseline regularization parameters we tried the range: $[100, 50, 10, 1, 0.1, 0.01, 0.05, 10^{-3}]$. To address the class-skew we assign a higher weight to the positives. For BSL-MTL, to tune the rank parameter 'k' we tried: [1, 5, 10, 25, 40, 60, 100] and the regularization parameter controlling the norm of U and V was tuned over the range λ=$\{1...10^{-3}\}$. For each task t, σ_t was varied over the values $[10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}]$ and μ_t was varied over $\{0, 0.25, 0.5, 0.75, 1\}$. The optimal setting was: $k = 10$.

2 10 fold cross validation

In the 10 fold cross-validation (CV) experiments, a much larger training set is available. So the single task baseline becomes much harder to beat as it can independently tune parameters for each task. All methods have a higher variance in their performance (we think this is due to the smaller and hence more variable test data in a 10 fold CV setting). Our method improves on only the *Hep-C* task.

	10 fold CV		
	Ebola	*Hep-C*	*Influenza*
STL	0.28±.11	0.74±.05	0.66±.08
MMTL	0.28±.11	0.68±.04	0.40±.03
Norm	0.20±.09	0.70±.06	0.44±.05
Rank	0.18±.08	0.70±.06	0.45±.04
MTPL	0.27±.08	0.67±.05	0.48±.06
IMC	0.16±.08	0.72±.06	0.45±.06
BSL-MTL	0.28±.11	0.82±.05	0.62±.02

Table 1: Area Under the Precision-Recall curve for 10 fold CV. The first row is the only single-task method and all others are multitask models.

3 Homologous interactions

Since we use data from several strains for each task, the PPI data contains some interactions that are interologs. We observed this for two of the tasks: *Hepatitis-C* and *Influenza A*. Note that we did not find any interologs *across* tasks. These homologous interactions in the various strains are reported by

different experimental studies and we believe their presence suggests the confidence of such interactions. Since we use only PPI derived from experimental methods (and not electronically inferred), we do not exclude any homologous interactions from our training.

Removing homologs from the evaluation:
The number of distinct PPI for which interologs or homologs exist is very small (≈20), but there are several homologs for each such PPI. Note that any benefits achieved from the presence of homologous interactions will be available to all the methods (since the models were built using identical train/test data). Here we present results on the 10% setting by removing all homologs from the test data (note: the *Ebola* task did not have any homologs). For lack of space we only mention the trend for our method (BSL-MTL) - *Hep-C*: 0.85 and *Influenza:* 0.45. The other methods show a similar trend; our method continues to outperform by significant margins.

4 Biological significance of interactions

Phosphorylation sites: We found the frequent occurrence of S and T and sometimes Y in the motifs striking and suspected this may be related to the amino acids being frequent targets of phosphorylation. Phosphorylated sites often serve as PPI sites, and databases such as Phosphosite [1] are repositories for known sites in human proteins. Since these are sites in human proteins, we searched for the patterns from the 4-mer motif in Fig 7 and found several to be flanking known phosphorylation sites in human proteins: `LLLs`, `LLLt`, `ILLs`, `PPPs`, `PIPs`, `PIPt`, `LIPs`, `PLLt` (lower-case indicates the putative phosphorylation site). This observation also supports the notion that the motifs predictive of interaction are biologically significant.

Evidence in IEDB[1]: We found experimental evidence for the significance for the virus motifs in the Immune Epitope[2] database (IEDB) [2]. The pattern `IVGG` from the *Hepatitis-C* motif in Fig. 5 is found in 53 epitopes. From the *Ebola* motifs in Fig 1, we find that `TLAT` is part of six different epitopes, `SLTT` appears in three epitopes. `PLIK`, `SLLL` from the *Influenza* motif are also found in many epitopes. Finding that the virus k-mer patterns predicted by our method are recognized by human antibodies is a further validation of its performance. Further, using higher dimensional k-mers (where k=7, 8, 9) as features in our model will give motifs from which complete epitopes can be derived. Our model thus has applications in epitope prediction as well, where conventional methods consist of scanning all possible k-mers from protein sequences to identify likely epitopes.

[1] `www.iedb.org`

[2] An epitope is a very short sequence from the virus that binds to human antibodies

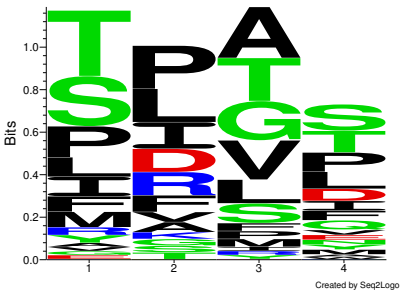

Figure 1: Sequence motif from top four-mer features specific to *Ebola* proteins

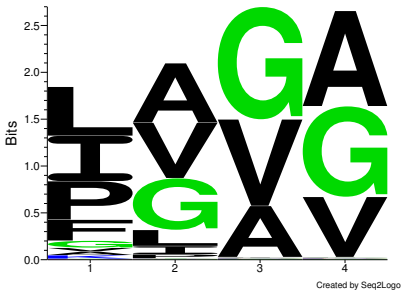

Figure 2: Sequence motifs specific to *Hepatitic-C* proteins that are also important to interactions.

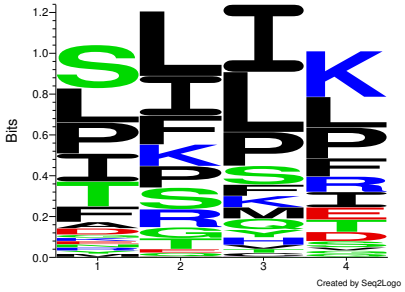

Figure 3: Sequence motifs specific to *Influenza* proteins that are also important to interactions.

5 Tri-mers

Below, we show the sequence motifs from the tri-mers found to be highly relevant to predicting interactions between human and viral proteins.

References

[1] Peter V Hornbeck, Bin Zhang, Beth Murray, Jon M Kornhauser, Vaughan Latham, and Elzbieta Skrzypek. Phosphositeplus, 2014: mutations, ptms and recalibrations. *Nucleic acids research*, 43(D1):D512–D520, 2015.

[2] Randi Vita, James A Overton, Jason A Greenbaum, et al. The immune epitope database (iedb) 3.0. *Nucleic acids research*, 43(D1):D405–D412, 2015.

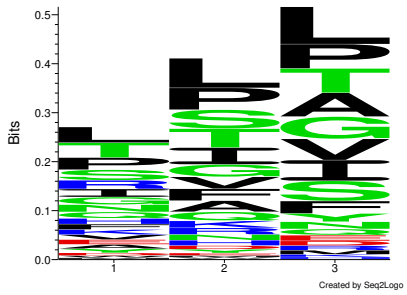

Figure 4: Sequence motifs specific to *Ebola* proteins that are also important to interactions.

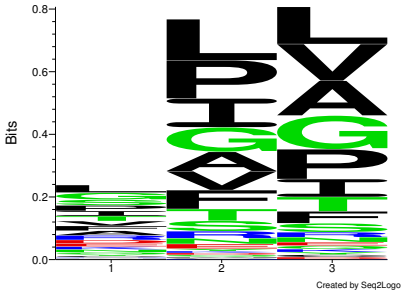

Figure 5: Sequence motifs specific to *Hepatitic-C* proteins that are also important to interactions.

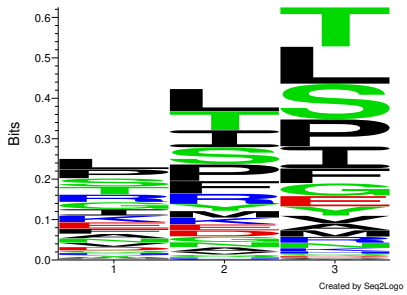

Figure 6: Sequence motifs specific to *Influenza* proteins that are also important to interactions.

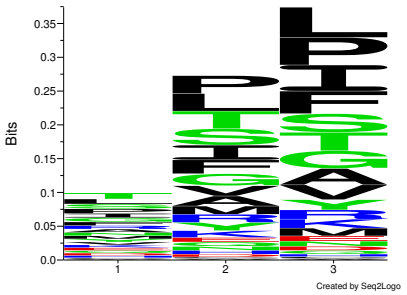

Figure 7: Sequence motif constructed from the top tri-mer features of human proteins